Week 3 Sponsor Meeting

Team ADS

Updates

- Amend project proposal, meeting minutes and wiki
- Domain summary
- Systematically extract info from URL…
  - 1. Proxy approach
  - 2. Pattern approach

Domain Summary

1. Create a text file for each domain, put all requests to that domain inside the file
2. Link each domain to one of 80+ databases.
3. Define user sessions.
4. Get proportion of post request and number of user sessions for each domain.

Extracting Info (Methodology)

❖ Proxy Approach

- Sort out search result page and download only.
- Use keywords (See below) to identify query results. Use file size to identify download.
  - Find the proper threshold
  - * Hypothesis: requests to full content results in large response size. When web assets have been eliminated, just by size, we can infer whether it's a download by size.

Query keywords:

"\Wquery=",
"\Wq=",
"\WsearchQuery="
"\Winput=" ((heinonline.org)
"\WqueryStr=" (lawnet.sg)

Rubbish keywords:

".gif",
".ico",
".js",

Extracting Info (Methodology)

❖ Proxy Approach

- Limitations:
  - Accuracy of keyword or size approach is unknown and hard to ascertain (unless manually check, but still, many page not accessible)
  - Cannot capture "read online" (i.e. viewing from embedded pdf viewer). May have info from url

Extracting Info (Methodology)

❖ Proxy Approach – Queries

- Example

https://**www.lawnet.sg**:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_WAR_lawnet3legalresearchportlet&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_action=openContentPage&contentDocID=/SLR/[2000]%203%20SLR(R)%200530.xml&**queryStr**=(Genelabs%20Diagnostics%20v%20Institut%20Pasteur)

Extracting Info (Methodology)

❖ Proxy Approach - Queries

- Example

https://**www.lawnet.sg**:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_WAR_lawnet3legalresearchportlet&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_count=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_action=openContentPage&contentDocID=/SLR/[2000]%203%20SLR(R)%200530.xml&**queryStr**=(Genelabs%20Diagnostics%20v%20Institut%20Pasteur)

Extracting Info (Methodology)

❖ Finding 1

- Total domains: 170
- Domains With query keywords 48

Extracting Info (Methodology)

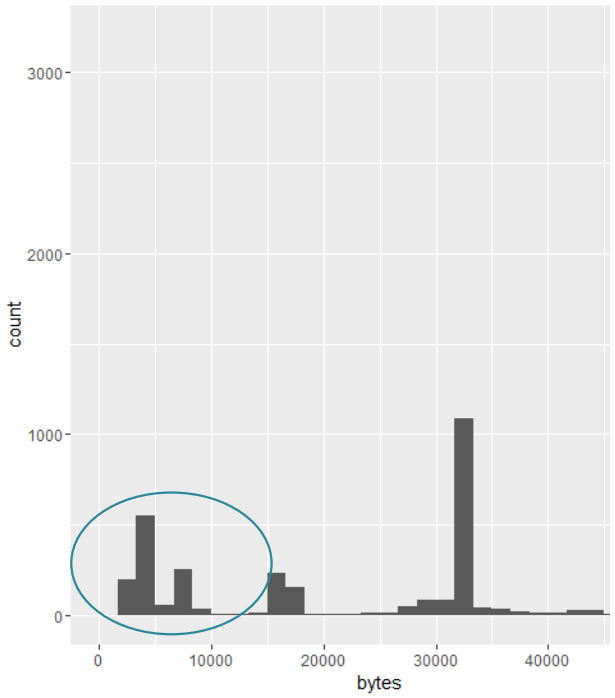
❖ Proxy Approach - Downloads

- Maybe other download format but “PDF” for now
- To find a proper size threshold, look at the distribution of requests containing keyword ".pdf" vs not containing keyword but eliminated web assets.

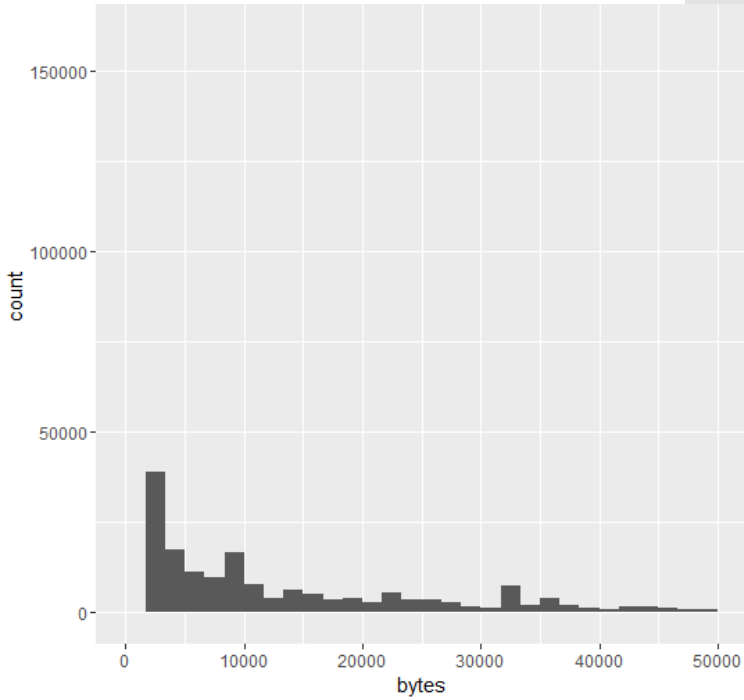
Extracting Info (Findings)

❖ Proxy Approach - Downloads

- Distributions roughly match hypothesis, except the "corner"
-



PDF

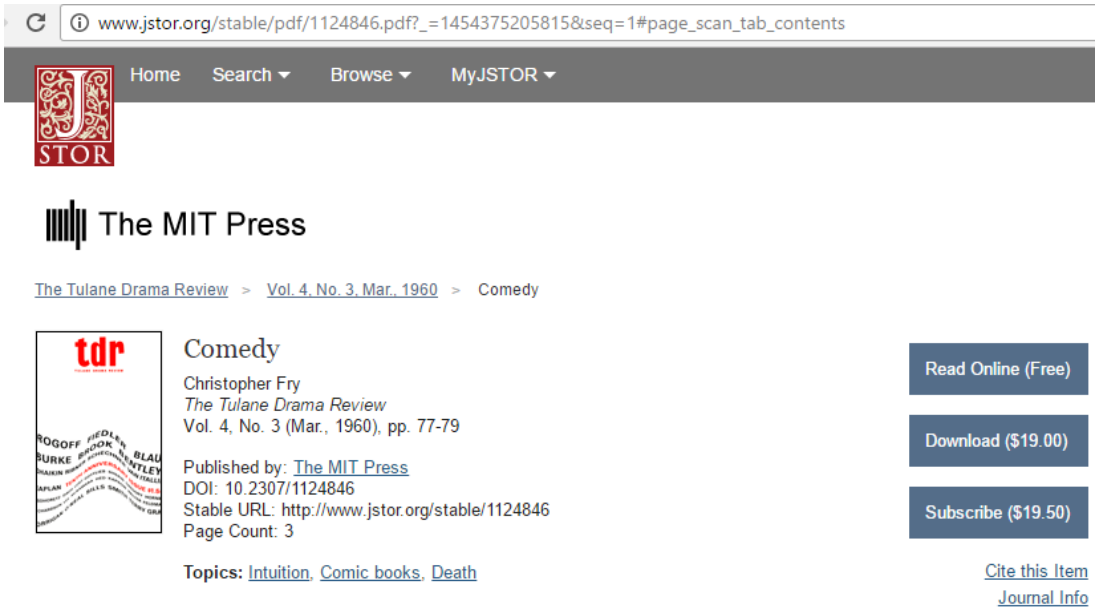


Non-PDF

Extracting Info (findings)

❖ Proxy Approach - Downloads

- Distributions roughly match hypothesis, except the "corner"
- Eg 1



Extracting Info (findings)

❖ Proxy Approach - Downloads

- Distributions roughly match hypothesis, except the "corner"
- Eg 2

http://delivery.acm.org:80/10.1145/2640000/2635922/p155-ray.pdf?ip=202.161.43.77&id=2635922&acc=ACTIVE%20SERVICE&key=FF6731C4D3E3CFFF%2E39D185EE56A58666%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=74373374o&CFTOKEN=43141124&__acm__=1454394669_ofccae66d0250484 4ce3e2d8d64b64f6
link

They are not real downloads

Extracting Info (Findings)

❖ Proxy Approach - Downloads

- Problem:
  - Sites use file name in URL, which is a ID and does not tell us anything (it's not a DOI)
- Analysis will be done on the proportion of "pdf" url with a real name

Extracting Info (Methodology)

❖ Pattern Approach

- Systematically find the useful url pattern and manually determine the meaning.
- Steps:

1. For each domain, find url patterns for real user requests. Starting from the most popular domains
2. For each pattern, find one example. Describe the user action for that example (whether it's a show search result, view item or download item or other).
3. For each pattern, get request counts and user counts.
4. For patterns that indicate user input, extract the input

Extracting Info (Methodology)

❖ Pattern Approach

- Inspired by:
  - Koppula et al, 2010, Learning URL Patterns for Webpage De-duplication, Retrieved from http://www.wsdm-conference.org/2010/proceedings/docs/p381.pdf
- To tokenize
  - sub-directories
  - Parameters

Extracting Info (Methodology)

❖ Pattern Approach

- Advantages:
  - Reduce human error
  - Generic
  - Accurate

Extracting Info (Methodology)

❖ Pattern Approach

https://www.lawnet.sg:443/lawnet/group/lawnet/result-page?p_p_id=legalresearchresultpage_WA
wnet3legalresearchportlet_action=searchBySearchTrailSearchId&_legalresearchresultpage_WAR_l
portlet_searchId=3700300
https://www.lawnet.sg:443/lawnet/group/lawnet/result-page?p_p_id=legalresearchresultpage_WA
wnet3legalresearchportlet_action=searchBySearchTrailSearchId&_legalresearchresultpage_WAR_l
portlet_searchId=3700801
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_
mn-2&p_p_col_count=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_action=openC
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_
=viewPDFSourceDocument&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_count=1&
1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_loadPage=0&_legalresearchpagecc
searchportlet_nextPage=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_viewType
M.xml&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_queryStr=%28%22%5B2004%5D+1
FileName=[2004]%201%20AC%20919.pdf&pdfFileUri=/English/70670-E-M/resource/[2004]%201%20AC%2
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_
mn-2&p_p_col_count=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_action=openC
20Cher%20Teck)
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_
=viewPDFSourceDocument&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_count=1&
+0367.xml&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_loadPage=0&_legalresear
3legalresearchportlet_nextPage=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_
010%5D+1+SLR+0367.xml&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_queryStr=%2
licitModel=true&pdfFileName=[2010]%201%20SLR%200367.pdf&pdfFileUri=/SLR/[2010]%201%20SLR%20
https://www.lawnet.sg:443/lawnet/group/lawnet/result-page?p_p_id=legalresearchresultpage_WA
wnet3legalresearchportlet_action=searchBySearchTrailSearchId&_legalresearchresultpage_WAR_l
portlet_searchId=3700488

Extracting Info (Methodology)

❖ Pattern Approach

- https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=legalresearchpagecontent_WAR_lawnet3legalresearchportlet&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_resource_id=viewPDFSourceDocument&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_count=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_documentID=%2FSLR%2F%5B2010%5D+1+SLR+0367.xml&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_loadPage=0&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_prevPage=-1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_nextPage=1&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_viewType=&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_contentDocID=%2FSLR%2F%5B2010%5D+1+SLR+0367.xml&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_queryStr=%28Goh+Suan+Hee+v+Teo+Cher+Teck%29&_legalresearchpagecontent_WAR_lawnet3legalresearchportlet_implicitModel=true&pdfFileName=[2010]%201%20SLR%200367.pdf&pdfFileUri=/SLR/[2010]%201%20SLR%200367/resource/[2010]%201%20SLR%200367.pdf

Pattern	count
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=XXX&p_p_lifecycle=D&XXX&p_p_mode=XXX&p_p_resource_id=XX&p_p_cacheability=XXX&p_p_col_id=XXX&p_p_col_count=D	20
…	…

Extracting Info (Methodology)

❖ Pattern Approach

Pattern	count
https://www.lawnet.sg:443/lawnet/group/lawnet/page-content?p_p_id=XXX&p_p_lifecycle=D&XXX&p_p_mode=XXX&p_p_resource_id=XX&p_p_cacheability=XXX&p_p_col_id=XXX&p_p_col_count=D	20
…	…